Scientific advance and theory integration in working memory

This is a prepublication copy of a manuscript published as:

Logie. R.H. (2018). Scientific advance and theory integration in working memory: Commentary on Oberauer et al. (2018) Benchmarks for models of short term and working memory. Psychological Bulletin.

Scientific advance and theory integration in working memory: Commentary on Oberauer et al. (2018) Benchmarks for models of short term and working memory.

Robert H Logie

University of Edinburgh

Oberauer and colleagues report the results of an ambitious project to identify key or 'benchmark' empirical findings that have been associated with the concept of working memory. An impressive range of findings is presented with the stated intention that the selection is as a-theoretical as possible, while classifying each finding according to whether or not it should be given high, medium or low priority to be explained by a generalisable theory of working memory. The three levels of priority are allocated by a combination of ratings from a wide range of working memory researchers and a consensus judgement by the authors.

There are at least two broad ways in which a collection into one place of an integrated set of brief descriptions of such a wide range of empirical findings could be a useful resource for researchers who consider themselves to be studying working memory. One of these, as stated by the authors, is to help development of theory and, in particular computational modeling of the concept of working memory by focusing on a common set of agreed empirical phenomena that should be explained by any such model or theory. There is indeed considerable merit in this, because genuine advances in cognitive theory generally, not just working memory, suffer from a lack of focus on common paradigms and common research questions, as well as from a lack of a broad agreement on the definition of what, precisely, is being studied. Major advances in economics, physics, evolutionary biology, and genetics along with many other sciences arise from multiple researchers and teams of researchers worldwide all focusing on the same, or a very similar research problem, even if they approach the problem from different directions. Working memory theory could certainly benefit from this kind of mass action. However, for this to work, it would be essential that individual theorists do not simply try to reinterpret findings to fit their favored theoretical framework, or change a parameter of an existing computational model without questioning the assumptions from which the model was developed. By focusing on the 'A' category of well established phenomena that have been thoroughly researched and replicated, there is a serious risk that working memory theorists will have already considered many of these, and will already have arguments as to how their favorite theory can explain them. This then simply encourages the proliferation of theories, rather than seeking common theoretical ground. Also, by focusing on explanation, theory development can become overly post-hoc, and circular, rather than integrative and predictive.

A second way, not considered by the authors, in which an integrated, and systematic presentation of empirical findings could be useful is in highlighting findings that some researchers might not have considered, or of which they had been unaware. This could serve as the basis for innovation in developing new paradigms, new methodology, and making new discoveries about the nature and functioning of working memory. Moreover, we would want our theories to predict novel empirical results rather than only to provide post-hoc explanations for known results. If this paper is successful in encouraging working memory theory development to focus on the 'A' category findings that have already been thoroughly researched, the result could be an overly conservative approach, leading to an endless cycle of scientific stagnation rather than scientific advance. By focusing on explanations for existing, well-established findings, category B and category C phenomena would become consigned to the electronic file drawer, and this would be accompanied by a rapid reduction in studies that seek out new findings, or more integrative theories that make novel predictions. Placing Category C findings in an Appendix will serve to accelerate their demise. It is not uncommon for new empirical findings to be serendipitous, forcing a reexamination of current theory, and development of new sets of predictions to be tested. Thus to ensure that we encourage innovation and genuine advance in knowledge, there is a strong argument that theory development should focus on the 'B' category findings, coupled with serious consideration of the 'C' category when developing theories and models. This is where we are most likely to find the results that are inconvenient for existing theories, and that should be addressed in theory development while assessing whether they can be replicated. Within this paper, by definition, each C category phenomenon has been reported by perhaps just one group of researchers, possibly driven by one particular theoretical framework. It is all too easy for researchers who favor an alternative theoretical framework to dismiss category C findings as not relevant or of less interest than category A findings that might be more

tractable for existing theories. A good example of a potentially very important, but possibly inconvenient finding in the category C appendix is 13.4 ‘Short-Term Retention Without Measurable Neurally Active Representations’. This finding presents a major challenge to theories that assume a lack of observed activation relative to a control condition means that the region of interest is not involved in supporting task performance. It suggests that *maintaining* a representation in working memory (rather than encoding or retrieving) will not necessarily be evident in the BOLD signal. This kind of finding also highlights how crucial it is to consider behavioral data and cognitive theory when interpreting the results of brain imaging. It is not alone in the C category Appendix as a finding with important theoretical implications.

A further reason for taking C category phenomena out of an appendix and placing them front and center, is that they may have found themselves with their lowly status because few researchers have addressed these phenomena. There is a strong element of a popularity contest in allocating phenomena to the A category. A democratic approach might not be the best approach to science, even if it dominates citation and download counts (which have their own flaws as metrics of quality and importance). This simply reflects the number of people who work, or have worked on those phenomena, and of course, the more people who work on a phenomenon, the more conceptual and actual replications there will be, and the more citations there will be to the published reports of those phenomena. It is also a concern that each researcher who contributed to these ratings will have her or his own priorities based on a particular theoretical perspective. Findings that are common, and have already been addressed within multiple theoretical frameworks will receive many votes. Findings that are important for a specific theory but have not been considered within other theories will only receive votes from the research group associated with that theory. There may be many such

findings, each with small numbers of votes from different groups of researchers, hence they become consigned to the B or C category. Many of the A category items may not be those that need to be addressed because they have already been addressed extensively. So the survey results and the replication criteria that were used together to generate the A, B, and C category findings are not independent, and aggregation of the ratings may be extremely misleading. Being in the A category does not mean that these are the most important phenomena for theory development. It could be that they are the easiest to explore experimentally, and/or are studied by a large number of researchers who share theoretical assumptions. It could also be that because these findings are already well established, researchers who are new to the field feel obliged to study them, thereby further increasing the community of researchers studying particular phenomena. Developing new ideas, new paradigms, and new, integrative theory is risky when developing an academic career. Individual researchers or groups could, of course, continue to generate novel avenues, findings and theory, but research funding and publishing present substantial challenges when the research is not in line with the current 'zeitgeist', or dominant paradigm and assumptions. Publication of the categorization, rather than an uncategorized repository of findings, will exacerbate this bias towards studying well established findings with existing theory rather than making a new contribution to knowledge and understanding.

There are also what might be called the 'category D' findings that have not been included. The authors are explicit about excluding a range of studies that have addressed what might be called 'applications' of working memory, for example in the study of reasoning, mental arithmetic, and language development. This certainly makes the current exercise more manageable, given the very large number of published studies that purport to study working memory. Arguably, such studies are of less interest if the primary goal is to develop

computational models of working memory, rather than, say, a computational model of mental arithmetic. However, in the spirit of the question that motivated Baddeley and Hitch (1974), 'What is it for?', we might expect that any conceptual (verbal/boxes and arrows) theory, or computational model ought to be able to predict as well as simulate and explain patterns of findings from these studies of how working memory might be used in everyday cognition. Arguably, this would be a stronger test of a theory than a focus on more artificial tasks that generate many of the category A findings.

Within what I refer to as 'category D' are many neuropsychological findings that are excluded by default rather than on principle. Benchmark 13.2 considers preserved working memory in amnesia, but there is no consideration, even in the C category, of the cases of impaired working memory with intact long-term episodic and semantic memory. Most of these cases appear to have specific impairments of immediate, serial ordered verbal recall (e.g. Basso, Spinnler, Vallar, & Zanobio, 1982; Warrington & Shallice, 1969; see reviews in Vallar & Shallice, 1990), but some other brain damaged individuals show intact verbal immediate recall, but impaired memory for novel visual or spatial stimuli (e.g. Beschin, Cocchini, Della Sala, & Logie, 1997; Logie, Beschin, Della Sala, & Denis, 2005; for reviews see Della Sala & Logie, 1993; Logie, 1995; Logie & Della Sala, 2005). These kinds of findings from studies of selective impairments have made important contributions to the development of the multi-component theoretical framework of WM (e.g. Baddeley & Logie, 1999), yet tend not to be considered by researchers who develop alternative theories. It may be that researchers who work with these kinds of patients are not comfortable with the idea of prioritizing findings based on how well established they are or how many votes they attracted, and therefore chose not to contribute to the benchmark project. However, it is a great pity that the authors did not consider including such findings, not least because they

offer the intriguing possibility of simulating focal damage within computational models to explore whether the ‘damaged’ models generate the observed patterns of cognitive impairment and sparing in multiple single case studies of brain damaged individuals.

The argument thus far has focused on the issue of which of the three categories developed by Oberauer et al. should be given highest priority. However, while having a comprehensive collection of research findings could be an extremely useful resource, it is not clear that assigning any kind of differential priority labels to different findings helps our understanding of working memory or helps theory development. The decision to assign high, medium and low priority to each finding is, at first glance, entirely rationale. Findings that have been replicated multiple times, and that have been ranked highly by researchers who have chosen to contribute to these rankings are given the highest priority. However, as argued above, it may be the findings that are novel, or have been shown in only a small number of studies, or tend to be considered only within one theoretical framework that could be the most informative, and offer the greatest potential for substantial advances in working memory theory.

An alternative, and potentially much more useful approach could be to provide an uncategorised repository of findings, with the A, B, and C labels removed, along with encouragement for researchers to give serious consideration to the potential merits of impartial comparison of alternative theories. The selection of findings for this comparison could be driven by the theories that are to be compared, not by how widely known are the findings. For example, Theory 1 might be associated with a particular set of Findings X, whereas Theory 2 might be associated with a different set of Findings Y. Theory 1 would then be set the challenge of addressing Findings Y, and Theory 2 would attempt to address

Findings X. Findings X and Y would be selected from those considered important for Theories 1 and 2, not by the A,B,C categorization for generalised priority. There might be requirements to modify the paradigms commonly used by Theory A and Theory B for the purposes of this comparison, and each theory would be set the task of generating its own sets of predictions in advance, to avoid the temptation to develop post-hoc explanations for unexpected data patterns. It would be essential that researchers associated with each theory remain open-minded about the possibility that neither theory will predict the results. This opens the opportunity for a step change in theory development to emerge in a new, more integrated Theory 3, that can account for, and would have predicted both findings X and Y. This kind of theory comparison is very different from the common practice in cognitive psychology of making predictions from one theory, and designing experiments to test those predictions. Also common is to set up an alternative theory against which to assess a favored theory, but with a paradigm that most often gives some advantage to the favored theory.

The fact that genuinely impartial theory comparison is rare in cognitive psychology was noted over 30 years ago by Watkins (1984), who suggested that cognitive theories are like toothbrushes: fine for individual use, but we would rather not use one belonging to someone else. Gigerenzer (2010) developed this concern by noting that we teach our students how to test a theory, but we tend not to teach students how to construct a theory. Moreover, Gigerenzer claimed that often labels and descriptions of phenomena, or statements of dichotomies are presented as theories, so that we develop new labels for phenomena rather than making major new theoretical advances. Maybe our goal should be for a theory to act, not like a toothbrush, but more like recipes that we are happy to share with others, that can incorporate a range of well established and new empirical ingredients, and through impartial

co-operation between competing chefs may be integrated to develop new dishes from both familiar and newly discovered ingredients.

This approach will most likely require abandoning some long-held assumptions, and the outcome also might conflict with powerful introspections about our conscious experience of our own working memory in action. A long time ago Pylyshyn (1973) noted that not all conscious experience is necessarily functional in cognition, nor is all that is functional in cognition available to consciousness (for further discussion see Logie, 2016; Logie & Cowan, 2015).

The authors have a potentially, extremely valuable resource, which could become even more valuable over time, but giving the highest priority to the most established findings, and excluding or diminishing the status of less widely reported findings could hinder, rather than facilitate theory development. Moreover, the text would have benefitted from cross referencing between benchmarks that would have helped readers to navigate among the various findings within and between categories. For example, is there any consistency between findings from neuroscience and neuropsychology, and behavioural findings regarding the effects of distracters in serial ordered recall? An integrative theme throughout that highlights consistencies and inconsistencies across different methodologies as well as across different paradigms would be much more helpful to the working memory community. Setting a major research agenda based on high priority for well established findings inevitably sidelines the new, but less well established findings that might be just as, if not more important for theory development than those that have been studied extensively.

In sum, I am not convinced that the aim of attempting to identify benchmark findings in working memory and the imposition of differential ratings of importance for each finding is of benefit to the working memory research community, and may be a formula for slowing progress theoretically and empirically, by stifling innovation in developments of methodology or in the generation of new findings that are crucial elements for advancing knowledge. However, if the benchmark theme were to be dropped, then the collation in a regularly updated and comprehensive web-based repository of cross-referenced published findings that have been subject to rigorous peer review could be a very valuable resource. Such a repository would help avoid 'reinvention of the wheel' when researchers report as new, a finding that was reported some time ago. It could be dynamic not only by adding new findings, but also by deleting or 'flagging' findings that suffer from multiple failures to replicate. Most important, it would act as a resource for developing, new, more integrated rather than fractionated theory.

References

Baddeley, A.D. & Logie, R.H. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (eds.) *Models of Working Memory*, pp 28-61. New York: Cambridge University Press.

Basso, A., Spinnler, H. Vallar, G. & Zanobio, E. (1982). Left hemisphere damage and selective impairment of auditory verbal short-term memory: A case study. *Neuropsychologia*, 20, 263-274.

Beschin, N., Cocchini, G., Della Sala, S. & Logie, R.H. (1997). What the eyes perceive, the brain ignores: A case of pure unilateral representational neglect. *Cortex,* 33, 3-26.

Della Sala, S., & Logie, R.H. (1993) When working memory does not work. The role of working memory in neuropsychology. *In* F. Boller & H. Spinnler (Eds.). *Handbook of Neuropsychology Vol* 8, pp 1-63. Amsterdam: Elsevier Publishers BV.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory and Psychology*, 20, 733-743.

Logie, R.H. (2016). Retiring the Central Executive. *Quarterly Journal of Experimental Psychology***,** 69, 2093–2109.

Logie, R.H., Beschin, N., Della Sala, S. & Denis, M. (2005). Dissociating mental transformations and visuo-spatial storage in working memory. Evidence from representational neglect. *Memory*, 13, 430-434.

Logie, R.H. & Cowan, N. (2015). Perspectives on working memory. *Memory and Cognition*, 43, 315–324.

Logie, R.H. & Della Sala, S. (2005). Disorders of visuo-spatial working memory. In P. Shah and A. Miyake (Eds.) *Handbook of Visuospatial Thinking.* Cambridge University Press: New York, pp 81-120.

Pylyshyn, Z.W. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, 80, 1-24.

Vallar, G. & Shallice, T. (1990*). Neuropsychological Impairments of Short-Term Memory*. Cambridge: Cambridge University Press.

Warrington, E.K. and Shallice, T. (1969) The selective impairment of auditory verbal short-term memory. *Brain,* 92, 885-896.

Watkins, M. J. (1984). Models as toothbrushes. *Behavioral & Brain Sciences, 7,* 86.